MET O 11 TECHNICAL NOTE NO 95

Regression on non-orthogonal analytical base functions, orthogonal analytical base functions, and principal components.

R DIXON

1. Introduction

Much of the utility of principal components rests upon the theorem which states that a set of ordered eigenvectors will be more efficient on average at extracting variance from a dependent data set than any other set of linear functions of the data. Met O 11 Technical Note No 93 presents experimental evidence which, even though it relates to independent data sets, suggests that it may be desirable to study the conditions of validity and practical range of application for this theorem. Principal components have attracted wide interest since they were introduced into Meteorology in the fifties, but at a fundamental theoretical level there is a grey area containing questions which have remained largely undiscussed. This Note raises one such question, the assumptions as to the nature of the residuals when principal components are used in regression.

N.B. This paper has not been published. Permission to quote from it must be obtained from the Assistant Director of the above Meteorological Office Branch.

2. Estimation using non-orthogonal analytical base functions

Let there be a data vector $\underline{h} = \underline{h}(h_1, h_2, ----, h_m)$ and a set of analytical base functions $F_1, F_2, ----, F_n$. These analytical base functions can then each be evaluated over m discrete points of a physical domain to form the m x 1 base vectors $\underline{f}_1, \underline{f}_2, \underline{f}_3, ---- \underline{f}_n$. It may then be supposed that the "true" relationship between the data and the base vectors is given by

$$\underset{m1}{\underline{h}} = \underset{mn}{F} \cdot \underset{n1}{\underline{a}} + \underset{m1}{\underline{r}} \qquad (1)$$

where $\underset{mn}{F}$ is the matrix of column vectors

$$\underset{mn}{F} = (\underset{m1}{\underline{f}_1} \,|\, \underset{m1}{\underline{f}_2} \,|\, ---- \,|\, \underset{m1}{\underline{f}_n}) \qquad (2)$$

$\underline{a}$ is the coefficient vector

$\underline{r}$ is the vector of residuals which arises because we are considering the case where $m > n$ and so the functions cannot fit the data exactly.

The belief that (1) is the true structural model for the data may be based on physical reasoning, on some previous experimental evidence, or it may simply be postulated faute de mieux. Whatever the reason the problem arises of estimating $\underline{a}$, for $\underline{a}$ itself is generally unknown. A well-known and much used estimate is the ordinary least squares estimate (OLS) $\hat{\underline{a}}$ given by

$$\underset{n1}{\hat{\underline{a}}} = \underset{nm}{(\tilde{F}} \cdot \underset{mn}{F)^{-1}} \cdot \underset{nm}{\tilde{F}} \cdot \underset{m1}{\underline{h}} \qquad (3)$$

If the residuals in (1) are such that, in the population

$$E(\underline{r}) = \underline{0} \qquad (4)$$

and

$$E(\underset{m1}{\underline{r}} \underset{1m}{\hat{\underline{r}}}) = \sigma^2 \underset{mm}{I} \qquad (5)$$

where I is the unit diagonal matrix, E is the expectation operator, and σ is a scalar, then

$$E(\hat{\underline{a}}) = \underline{a} \qquad (6)$$

and

$$E(\hat{\underline{a}} - \underline{a})(\hat{\underline{a}} - \underline{a})^{\sim} \text{ is a minimum} \qquad (7)$$

(6) states that $\hat{\underline{a}}$ is unbiassed, whilst (7) states that the sampling variance - co-variance of the difference between the estimate $\hat{\underline{a}}$ and the true unknown $\underline{a}$ is a minimum amongst all such linear unbiassed estimators. Clearly (6) and (7) are very desirable properties for an estimate of $\underline{a}$ for they minimize the risk that one has a bad estimate.

If the residuals in (1) are such that (4) holds but in place of (5) we have

$$E(\underset{m1}{\underline{r}}\ \underset{1m}{\hat{\underline{r}}}) = \sigma^2 \underset{mm}{\Omega} \tag{8}$$

where $\underset{mm}{\Omega}$ is a more general matrix than I then the OLS estimate $\hat{\underline{a}}$ given by (3) does not have the property (7). The solution vector which does have the desirable minimum sampling variance property (7) I will denote by $\overset{*}{\underline{a}}$ and it is given by

$$\underset{n1}{\overset{*}{\underline{a}}} = (\underset{nm}{\tilde{F}} \cdot \underset{mm}{\Omega^{-1}} \cdot \underset{mn}{F})^{-1} \cdot \underset{nm}{\tilde{F}} \cdot \underset{mm}{\Omega^{-1}} \cdot \underset{m1}{\underline{h}} \tag{9}$$

Under the condition (8) it is the solution vector $\overset{*}{\underline{a}}$ which has the property

$$E(\overset{*}{\underline{a}} - \underline{a})(\overset{*}{\underline{a}} - \underline{a})^{\sim} \text{ is a minimum} \tag{10}$$

$\overset{*}{\underline{a}}$ is the fact the least-squares solution to a transformed version of the "true" relationship (1). The transformation is

$$\Omega^{-\frac{1}{2}} \cdot \underline{h} = \Omega^{-\frac{1}{2}} \cdot F \cdot \underline{a} + \Omega^{-\frac{1}{2}} \cdot \underline{r} \tag{11}$$

Thus if it is known, assumed, inferred or believed that (8) holds then the least-squares fitting procedure must be applied to solve (11) not (1) if the minimum variance solution is to be otained. This transformation of (1) by taking the scalar product of the square-root of Ω^{-1} through (1) is known as prewhitening, and the whole process of finding $\overset{*}{\underline{a}}$ is referred to as generalized least-squares(GLS) and $\overset{*}{\underline{a}}$ is known as the GLS estimate of $\underline{a}$.

It should be noticed that getting the best solution vector is dependent upon making a correct assumption as to the nature of the residuals. Furthermore, it is the residuals in (1), the "true" model equation, about which assumptions have to be made, not the residuals which are obtained as a result of the fitting, i.e. as the result of using $\hat{\underline{a}}$ or $\overset{*}{\underline{a}}$. With laboratory data it may be that a scientist can be reasonably sure of his assumptions because the data have been produced by a well understood and controlled process, but with meteorological data often produced by processes which are only imperfectly understood and over which we have no control the difficulty is a very real one. The residuals from the fitted process have to be used somehow to infer the validity or otherwise of the assumptions on the true residuals. That is to say for example that if it is assumed that (1), (4), and (5) hold then the residuals $\hat{\underline{r}}$ in

$$\underset{m1}{\underline{h}} = \underset{mn}{F} \cdot \underset{n1}{\hat{\underline{a}}} + \underset{m1}{\hat{\underline{r}}} \tag{12}$$

have to be used to test the validity of the assumptions (4) and (5) about the unknown residuals $\underset{m1}{r}$ in (1). It is something of a chicken and egg problem and it has a voluminous literature.

The OLS solution to (1) may be specified by the following two equations

$$\underset{m1}{h} = \underset{mn}{F} \cdot \underset{n1}{\hat{a}} + \underset{m1}{\hat{r}} \tag{13}$$

$$\underset{n1}{o} = \underset{nm}{\tilde{F}} \cdot \underset{m1}{\hat{r}} \tag{14}$$

This follows since, by virture of (14), taking $\tilde{F}$. through (13) leads in effect straight to the OLS solution (3). Equation (14) requires the residual vector $\hat{r}$ to be orthogonal to the columns of F. The fact that (13) and (14) together specify the OLS solution to (1) is of central importance to the theme of this Note, as will emerge in Section 4.

3. Estimation using orthogonal analytical base functions

The model (1) is now replaced by

$$\underset{m1}{h} = \underset{mn}{\Phi} \cdot \underset{n1}{b} + \underset{m1}{r} \tag{15}$$

where $\underset{mn}{\Phi}$ is the matrix of orthonormal column vectors

$$\underset{mn}{\Phi} = \left(\underset{m1}{\phi_1} \,|\, \underset{m1}{\phi_2} \,|\, \underset{m1}{\phi_3} \,|\, \text{----} \,|\, \underset{m1}{\phi_n} \right) \tag{16}$$

the $\underline{\phi}$-vectors being orthonormal in the simple sense that

$$\underset{nm}{\tilde{\Phi}} \cdot \underset{mn}{\Phi} = \underset{nn}{I} \tag{17}$$

The OLS estimate for $\underline{b}$ is then

$$\underset{n1}{\hat{b}} = \underset{nm}{\tilde{\Phi}} \cdot \underset{m1}{h} \tag{18}$$

which is computationally convenient in the sense that each coefficient in $\hat{b}$ is obtained by a simple scalar product

$$b_i = \underset{1m}{\tilde{\phi}_i} \cdot \underset{m1}{h} \qquad i = 1, 2, \text{---}, n \tag{19}$$

The use of orthogonal base vectors has done nothing to change the considerations as to the nature of the residuals. The arguments of the previous Section still apply and $\hat{b}$ will be the unbiassed minimum variance estimate of $\underline{b}$ if and only if (4) and (5) hold. Indeed in most applications (1) and (15) will be the same model because the $\underline{\Phi}$ - matrix will be obtained from the F matrix by some factorization process such as the Gram-Schmidt or Householder algorithms so that

$$\underset{m\,n}{F} = \underset{m\,n}{\underline{\Phi}} \cdot \underset{n\,n}{C} \tag{20}$$

and $\underline{b}$ and $\underline{a}$, and $\hat{\underline{b}}$ and $\hat{\underline{a}}$ will be related by the simple linear transformation

$$C \cdot \underline{a} = \underline{b} \quad , \quad C \cdot \hat{\underline{a}} = \hat{\underline{b}} \tag{21}$$

However the use of orthogonal base vectors does give rise to a useful artifice. We can generate a full set of m $\underline{\phi}$ - vectors to form the m x m matrix

$$\underset{m\,m}{\underline{\Phi}} = \left(\underset{m\,1}{\underline{\phi}_1} \,\vdots\, \underset{m\,1}{\underline{\phi}_2} \,\vdots\, \underset{m\,1}{\underline{\phi}_3} \,\vdots\, \cdots\cdots \,\vdots\, \underset{m\,1}{\underline{\phi}_m} \right) \tag{22}$$

with the $\underline{\phi}$ - vectors orthonormal in the simple sense that

$$\underset{m\,m}{\underline{\tilde{\Phi}}} \cdot \underset{m\,m}{\underline{\Phi}} = \underset{m\,m}{\underline{I}} \tag{23}$$

The data vector $\underline{h}$ can now be completely represented in terms of these $\underline{\phi}_s$ as

$$\underset{m\,1}{\underline{h}} = b_1 \underset{m\,1}{\underline{\phi}_1} + b_2 \underset{m\,1}{\underline{\phi}_2} + \cdots\cdots + b_m \underset{m\,1}{\underline{\phi}_m} \tag{24}$$

There is no residual vector as the data is fitted exactly. We can now suppose that the terms in (24) have been ordered according to some principle and we can then decide to discard all but n of these terms and represent the data as

$$\underset{m\,1}{\underline{h}} = b_1 \underset{m\,1}{\underline{\phi}_1} + b_2 \underset{m\,1}{\underline{\phi}_2} + \cdots\cdots + b_n \underset{m\,1}{\underline{\phi}_n} \tag{25}$$

This is quite legitimate mathematically as the orthogonality of the $\underline{\phi}_s$ makes their individual contributions independent of each other. Furthermore the ordering and selection principle which has decided the terms to be retained in (25) may have some apparently quite sound physical reasoning behind it and we may feel justified in arguing that the discarded (m - n) terms can contribute little but noise. Nevertheless, this artifice needs to be examined more closely.

Although the representation of $\underline{h}$ in terms of n vectors is often written as in (25) above, yet (25) is not strictly correct. Since $n < m$ the terms can no longer fit the data exactly and a vector of residuals must be introduced. Thus strictly speaking (25) has to be written as

$$\underset{m\,1}{\underline{h}} = b_1 \underset{m\,1}{\underline{\phi}_1} + b_2 \underset{m\,1}{\underline{\phi}_2} + \cdots\cdots + b_n \underset{m\,1}{\underline{\phi}_n} + \underset{m\,1}{\underline{r}} \tag{26}$$

where the use of the different diacritic Ψ over the $\underline{r}$ merely indicates an absence of any commitment, at this stage, as to the nature of the $\underline{r}$ in (26). But the nature of the $\underline{r}$ in (26) can in fact be elicited, for if the representation (24) fits the data exactly whilst the representation (26) fits it leaving a residual $\underline{r}$ it follows that this residual vector $\underline{r}$ is itself represented by the discarded $\underline{\phi}$ terms and that

$$\underset{m1}{\underline{r}} = b_{n+1} \underset{m1}{\underline{\phi}_{n+1}} + b_{n+2} \underset{m1}{\underline{\phi}_{n+2}} + \cdots + b_m \underset{m1}{\underline{\phi}_m} \qquad (27)$$

Now, since all the $\underline{\phi}$-vectors in (24) are orthogonal to each other, all the $\underline{\phi}$-vectors in (27) are orthogonal to all the $\underline{\phi}$-vectors in (26). This means that $\underline{r}$ itself in (26) is orthogonal to all the $\underline{\phi}$-vectors in (26). Thus if we express this fact, and (26) itself, in matrix-vector form we have

$$\underset{m1}{\underline{h}} = \underset{mn}{\underline{\Phi}} \cdot \underset{n1}{\underline{b}} + \underset{m1}{\underline{r}} \qquad (28)$$

$$\underset{n1}{\underline{0}} = \underset{nm}{\underline{\tilde{\Phi}}} \cdot \underset{m1}{\underline{r}} \qquad (29)$$

But these two equations have a familiar look and referring back it is seen that they are the analogues, for the orthogonal case, of equations (13) and (14). Equations (28) and (29) are in fact the two conditions which specify the OLS solution to (15). This follows because taking $\underline{\tilde{\Phi}}\cdot$ through (28) leads by virture of (29) and (17) straight to (18). Thus we may drop the non-committal diacritic Ψ and rewrite (28) and (29) with the appropriate diacritic as

$$\underset{m1}{\underline{h}} = \underset{mn}{\underline{\Phi}} \cdot \underset{n1}{\underline{\hat{b}}} + \underset{m1}{\underline{\hat{r}}} \qquad (30)$$

$$\underset{n1}{\underline{0}} = \underset{nm}{\underline{\tilde{\Phi}}} \cdot \underset{m1}{\underline{\hat{r}}} \qquad (31)$$

This Section brings out the fact that the artifice of selecting from a complete set of orthogonal base vectors, or ordering and truncating the set, and forming the coefficients of a data set representation by taking scalar products (19) is fully equivalent to the OLS process and implies the same assumptions on the residuals. This last point is crucial, for it is not obvious and may easily be overlooked.

4. Estimation using Principal Components

In this case instead of using base vectors derived by evaluating a set of analytical functions over the discrete points of the domain we derive the base vectors directly from a set of data vectors. Taking N data vectors $\underline{h}_1, \underline{h}_2, \underline{h}_3, \cdots, \underline{h}_N$ and viewing these N vectors as m x 1 column vectors we assemble them to form the dependent data matrix

$$\underset{mN}{H} = \left(\underset{m1}{\underline{h}_1} \,\vdots\, \underset{m1}{\underline{h}_2} \,\vdots\, \underset{m1}{\underline{h}_3} \,\vdots\, \cdots \,\vdots\, \underset{m1}{\underline{h}_N} \right) \qquad (32)$$

Now by taking the scalar product $\cdot\tilde{H}$ we can form the symmetric matrix C, where

$$\underset{mm}{NC} = \underset{mN}{H} \cdot \underset{Nm}{\tilde{H}} \tag{33}$$

The effect of forming the scalar product has been to eliminate the N-space. Also if the components of the individual $\underline{h}$ vectors have been measured from the individual means of each vector then C is the covariance matrix of the data. C can now be factored into

$$\underset{mm}{C} = \underset{mm}{U} \cdot \underset{mm}{\Lambda} \cdot \underset{mm}{\tilde{U}} \tag{34}$$

where

$$\underset{mm}{\Lambda} = \mathrm{diag}\,(\lambda_1, \lambda_2, \lambda_3, \text{----}, \lambda_m) \tag{35}$$

is a diagonal matrix of eigenvalues of $\frac{1}{N}H.\tilde{H}$ and

$$\underset{mm}{U} = \left(\underset{m1}{\underline{u}_1} \,|\, \underset{m1}{\underline{u}_2} \,|\, \underset{m1}{\underline{u}_3} \,|\, \text{----} \,|\, \underset{m1}{\underline{u}_m}\right) \tag{36}$$

is a modal matrix, the columns of which are the eigenvectors corresponding to $\lambda_1, \lambda_2, \lambda_3, \text{------}, \lambda_m$. Without loss of generality we can assume that $\lambda_1 > \lambda_2 > \lambda_3 > \text{----} > \lambda_m$ The modal matrix U has the property

$$\underset{mm}{\tilde{U}} \cdot \underset{mm}{U} = \underset{mm}{I} \tag{37}$$

Since the $\underline{u}_i$ are m-vectors they can be used as base-vectors for the respresentation of any other m-vector. In particular they may be used as base vectors for the representation of the individual vectors of the dependent data set, so that we have

$$\underset{m1}{\underline{h}_i} = b_{1i}\,\underset{m1}{\underline{u}_1} + b_{2i}\,\underset{m1}{\underline{u}_2} + b_{3i}\,\underset{m1}{\underline{u}_3} + \text{----} + b_{mi}\,\underset{m1}{\underline{u}_m} \qquad i = 1, 2, \text{---}, N \tag{38}$$

where the b_{ij} are sets of coefficients (not, of course, the same as the b_i in Section 3).

The m equations (38) may be rewritten in matrix-vector notation as

$$\underset{mN}{H} = \underset{mm}{U} \cdot \underset{mN}{B} \tag{39}$$

From (39) and (37) it follows that

$$\underset{mm}{\tilde{U}} \cdot \underset{mN}{H} = \underset{mN}{B} \tag{40}$$

and that therefore

$$\underset{Nm}{\hat{B}} = \underset{Nm}{\tilde{H}} \cdot \underset{mm}{U} \tag{41}$$

and consequently

$$\underset{mN}{B} \cdot \underset{Nm}{\hat{B}} = \underset{mm}{\tilde{U}} \cdot \underset{mN}{H} \cdot \underset{Nm}{\tilde{H}} \cdot \underset{mm}{U} \tag{42}$$

But from (33) this is

$$\underset{mN}{B} \cdot \underset{Nm}{\tilde{B}} = \underset{mm}{\tilde{U}} \cdot \underset{mm}{NC} \cdot \underset{mm}{U} \tag{43}$$

and now, using (34) and (37) again, this comes to

$$\underset{mN}{B} \cdot \underset{Nm}{\tilde{B}} = N \underset{mm}{\Lambda} \tag{44}$$

Λ is the diagonal matrix of eigenvalues and so it transpires that the row vectors of B are orthogonal.

These are the basic equations and relationships of principal component analysis and using them one can prove the main well-known theorem to the effect that such a set of ordered eigenvectors will be more efficient on average at extracting variance from the dependent data set than any other set of linear funtions of the data. Note the words "on average" in this statement of the theorem. It does not follow that the theorem is true for any particular individual data vector taken from the set.

In the light of this theorem it is natural to attempt to obtain an economical representation of the data vectors of the set by truncating (38) on some criterion. Thus dropping the subscript and taking $\underline{h}$ as a typical data vector we will have

$$\underset{m1}{\underline{h}} = b_1 \underset{m1}{\underline{u}_1} + b_2 \underset{m1}{\underline{u}_2} + \cdots + b_n \underset{m1}{\underline{u}_n} + \overset{*}{\underset{m1}{\underline{r}}} \tag{45}$$

or in matrix-vector form

$$\underset{m1}{\underline{h}} = \underset{mn}{U} \cdot \underset{n1}{\underline{b}} + \overset{*}{\underset{m1}{\underline{r}}} \tag{46}$$

the coefficients being found by taking the scalar products

$$b_j = \underset{1m}{\tilde{u}_j} \cdot \underset{m1}{h} \qquad (47)$$

Notice that since the n terms cannot fit the data exactly it has been necessary in (45) to introduce the unspecified residual vector $\underset{}{\hat{r}}$. In the general literature of principal components, including the meteorological literature, this residual vector goes virtually unacknowledged in the theoretical discussion. Here the main theoretical point to be made is that once placed plainly in evidence as in (45) or (46) it does not long remain unspecified. It is clear that $\underset{}{\hat{r}}$, being composed of all the discarded $\underset{}{u}$ terms, is orthogonal to each of the column vectors of $\underset{mn}{U}$ and we are in fact dealing with precisely the same situation as in the case of equation (26) in Section 3 on orthogonal analytical base functions. The argument following (26) applies equally in the case of (45) and leads in exactly the same way to

$$\underset{m1}{h} = \underset{mn}{U} \cdot \underset{n1}{\hat{b}} + \underset{m1}{\hat{r}} \qquad (48)$$

$$\underset{n1}{o} = \underset{nm}{\tilde{U}} \cdot \underset{m1}{\hat{r}} \qquad (49)$$

In other words the truncation of the complete set of eigevectors and the formation of the coefficients as scalar products (47) leads to the OLS solution. We have in fact found the OLS solution to a model

$$\underset{m1}{h} = \underset{mn}{U} \cdot \underset{n1}{b} + \underset{m1}{r} \qquad (50)$$

which has not been explicity postulated. The general argument and terminology of the principal component approach somewhat obscures the fact that a model has been postulated, but such is the case and the OLS solution $\underset{}{\hat{b}}$ found by the above procedure will be the desirable "best" (i.e. minimum variance) solution only if (5) applies, just as in the cases of the analytical base functions dealt with in Sections 2 and 3. In other words a principal component enthusiast has as much of an obligation to state and test his assumptions concerning the residuals as does an enthusiast for polynomials, fourier terms, exponentials, etc.

The principal components concept appears to have come over to meteorology from psychology and biology and in these sciences the basic theory, equations (32) to (44), ~~has been presented~~ has been presented with a subtle difference of emphasis. In the presentation as given in Section 4 the condition (44), revealing the orthogonality of the rows of the coefficient matrix B, emerges as an almost incidental consequence of the use of eigenvectors. In the presentation favoured by psychologists, biologists, and some meteorologists the condition (44) is made a central requirement. Indeed, it is sometimes referred to as the "principal component property". It is then shown that the eigenvectors of the sample covariance matrix are the only base vectors which permit this. Mathematically there is nothing to choose between the two approaches. The argument and terminology accompanying either of them tends to obscure the point made in this Note.

R Dixon

R Dixon
Met O 11

September 1977